



TETRA TECH EC, INC.

August 28, 2018

VIA ELECTRONIC MAIL AND OVERNIGHT DELIVERY

Lawrence Lansdale  
Environmental Director  
Base Realignment and Closure Program Management Office  
U.S. Department of the Navy  
33000 Nixie Way  
San Diego, CA 92147  
Lawrence.Lansdale@navy.mil

**Re: Hunters Point Naval Shipyard  
Tetra Tech EC, Inc. Technical Comments  
Draft Parcel G Removal Site Evaluation Work Plan (June 2018)**

Dear Mr. Lansdale:

Enclosed are Tetra Tech EC, Inc.'s ("TtEC's") Technical Comments on the June 2018 Draft Parcel G Removal Site Evaluation Work Plan ("Draft Work Plan").

### **Background**

TtEC has conducted remedial work at the Hunters Point Naval Shipyard Site ("HPNS") since 2004 pursuant to a series of contracts with the Navy. In 2012, the Navy and TtEC investigated a discrete set of soil samples that did not appear to be representative of the locations sampled. In 2014, TtEC issued a comprehensive report describing the investigative steps and corrective actions taken, all of which was done in close consultation with the Navy and accepted under the terms of the contract. Though no one admitted to wrongdoing at that time, two TtEC employees, who previously worked for the subcontractor New World Environmental ("NWE") were placed on leave following the investigation. Those two individuals, Justin Hubbard and Steven Rolfe, later admitted to the U.S. Department of Justice that they switched certain soil samples at issue in the investigation so that "clean" soil was analyzed rather than soil from the sampling locations. Hubbard and Rolfe were prosecuted and sentenced to prison, and TtEC fully supports the Government's actions in those cases.

Following the 2012 investigation and 2014 report, a handful of former NWE employees made accusations that other fraudulent sampling activities took place at HPNS. It is noteworthy that the accusers stand to benefit financially if their accusations are advanced by the Government. TtEC has investigated their allegations and determined that they are wholly unsubstantiated.

TtEC stands by its work at HPNS and has continued to cooperate with the Navy in meeting the objectives for the HPNS Site, while defending itself and its valued employees against these baseless accusations.

In 2017, the Navy contracted with CH2M Hill, Inc. and other competitors of TtEC to prepare a desk review of data collected by TtEC at HPNS. These consultants applied incorrect, and in many cases, arbitrary criteria to the HPNS data, using invalid statistical and analytical methods, resulting in the Draft Data Evaluation Reports (“Draft Reports”) that improperly question significant portions of TtEC’s work. Significantly, the Draft Reports make no attempt to align their analysis with the Statements of Work, Performance Measurements, or other terms of the contracts under which TtEC performed its work over a period of more than eight years. We respectfully submit that the Draft Reports fall short of the scientific and evidentiary standards applied to technical reports. Accordingly, it would be inappropriate for the Navy to rely on the Draft Reports as a basis for decision-making at HPNS.

### **The Draft Reports**

The Draft Reports include a Draft Building Radiation Survey Data Initial Evaluation Report (Mar. 2018); Draft Radiological Data Evaluation Findings Report for Parcel C Soil (Nov. 2017); Draft Radiological Data Evaluation Findings Report for Parcels D-2, UC-1, UC-2, and UC-3 Soil (Oct. 2017); Draft Radiological Data Evaluation Findings Report for Parcels B and G Soil (Sept. 2017); and Draft Radiological Data Evaluation Findings Report for Parcel E Soil (Dec. 2017). The Draft Reports were prepared by consultants hired by the Navy, including CH2M Hill, Inc. (“CH2M Hill”) (now a subsidiary of Jacobs Engineering Group, Inc.) and Battelle Memorial Institute (“Battelle”).

The Draft Reports rely on arbitrary logic tests, inappropriate statistical tests, and misleading graphics, all of which are misapplied and misinterpreted to reach incorrect conclusions. The misuse of the logic and statistical tests and the misleading graphics results in a large percentage of HPNS data being incorrectly identified as potentially suspect. In addition, the Draft Reports do not consider alternative scientific explanations for any potential data issues, such as the well documented highly variable soil conditions at HPNS or sensitivity to background radiation levels.

Specifically, the logic tests used in the Draft Reports impose a series of arbitrary requirements on the conditions under which samples are collected (*e.g.*, the relative timing of sample collection and analysis) and identify any deviation from those arbitrary requirements as evidence or “potential evidence” of “potential data manipulation.” However, the logic tests have no foundation in the contractual requirements for work at the Site, nor do they have any scientific or technical foundation. In the vast majority of cases, failure of the logic tests is easily explained by ordinary field sampling or laboratory operations, such as the re-analysis of samples according to the laboratory Standard Operating Procedures approved by the Navy, or by benign or innocent errors in data and information processing that are unavoidable in large projects. Failure to meet the requirements imposed by these arbitrary logic tests is not evidence of data irregularities.

The “statistical tests” used in the Draft Reports are flawed, and in some cases, they are not actually statistical tests at all. For example, the defects in the Draft Reports’ statistical analysis include the following:

- The Draft Reports’ use of the Kolmogorov-Smirnov (“KS”) test to identify “statistically different” populations of data ignores natural heterogeneity in soils at the Site and differences in conditions under which samples were collected.
- The application of Benford’s Law tests to data with an insufficiently wide range of values (*e.g.*, <sup>228</sup>Ac data) incorrectly identifies hundreds of data points as potentially suspect.
- The hierarchical “clustering” analysis performed by the consultants is not a statistical test at all, but rather, a subjective approach to data assessment.
- The confidence intervals in the Draft Reports are either computed incorrectly or based on arbitrary unstated assumptions. (*See, e.g.*, Appendix A of the Draft Parcel G Report, which purports to determine confidence intervals for SU0085 based on a single sample.)
- The Draft Reports do not provide any information about the procedures used to identify “outliers” in the dataset or to flag unusual data, and the large number of outliers identified in the Draft Reports strongly suggest the methodology generated results that are meaningless. Further, the flagging of outlying values is inconsistent, and is inappropriately based on a univariate as opposed to a multivariate analysis.

The misapplication of these statistical tests results in a large percentage of data being incorrectly identified as potentially suspect. Moreover, even where statistical differences may be present, the Draft Reports do not consider alternative scientific explanations for the differences, such as heterogeneity in soil conditions or sensitivity to background radiation levels. Indeed, as evidenced in the December 13, 2016 Scoping Meeting Summary attached as Attachment 1 to the August 2018 Draft Parcel G Removal Site Evaluation Sampling and Analysis Plan (“Draft SAP”), the statistical tests apparently were designed to arrive at the conclusion that TtEC engaged in “potential data falsification.” In response to concerns expressed by the U.S. Environmental Protection Agency (“EPA”) that the statistical tests would not return any evidence of data anomalies, the consultants provided assurance that the tests were designed to “identify anomalies in the data.” Moreover, when EPA “raised questions” about “what approach will be taken if data testing methods do not recommend sampling in places where allegations have pointed to,” one of the consultants explained that confirmation sampling would be done where former workers alleged wrongdoing. Thus, it is clear that the Draft Reports do not provide an objective analysis of the data collected at HPNS.

Finally, even if taken at face value, the Draft Reports only tentatively conclude that there is “potential evidence” of “potential data manipulation.” It is unclear what the term “potential evidence” means, but nonetheless the Draft Reports are being inappropriately relied on by the Navy and EPA to make critical decisions about HPNS.

## **The Draft Work Plan**

The Draft Work Plan compounds these errors by relying on the Draft Parcel B and G Report to call for extensive excavation and sampling in areas where the flawed Draft Report found evidence of “potential” data manipulation, as well as additional work in other areas of Parcel G where no evidence of “potential” data manipulation was found. The Draft Work Plan incorrectly and repetitively claims that “[a]n independent third-party evaluation of TtEC data found evidence of manipulation and falsification at Parcel G.” This statement, referencing the Draft Parcel B and G Report, is not true. First, the “independent third party” is not in fact independent, since CH2M Hill and Battelle have apparent organizational conflicts of interest. Second, the Draft Reports only purport to identify potential evidence of potential data manipulation, not actual evidence of data falsification. Finally, the approach taken in the Draft Reports to attempt to identify “potential” data manipulation is deeply flawed. Thus, the Navy has invested considerable time and funds on an analysis produced by self-interested consultants, which now potentially puts the Navy at risk of spending more time and money performing unnecessary and duplicative work on Parcel G.

The Draft Work Plan does not satisfy the Navy’s obligation to engage in reasoned decision making, and any action to implement the Work Plan as drafted would violate the Comprehensive Environmental Response, Compensation and Liability Act of 1980, the Base Realignment and Closure Acts of 1988 and 1990, and the Defense Environmental Restoration Program, Chapter 160 of Title 10, United States Code.

## **Consultants’ Conflicts of Interest**

CH2M Hill and Battelle were the primary consultants that conducted the desk review of data collected by TtEC at HPNS, including at Parcel G. CH2M Hill and Battelle authored the Draft Reports for the Navy.

CH2M Hill is also the primary consultant that will be performing the work described in the Work Plan, and Battelle is the primary consultant that will supervise the work performed by CH2M Hill. Despite CH2M Hill’s and Battelle’s central role in preparing the Draft Reports, the Work Plan characterizes the Draft Reports as the work of an “independent third party.” This statement is misleading, and fails to acknowledge the conflict of interest created by CH2M Hill and Battelle, on the one hand evaluating TtEC’s work at the Site, and on the other hand benefiting financially from the work that these purportedly independent Draft Reports will generate.

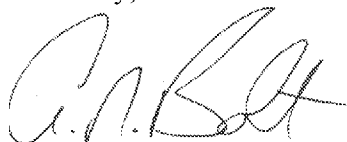
The Federal Acquisition Regulation (“FAR”) requires that Government contracting officers prevent “the existence of conflicting roles that might bias a contractor’s judgment.” FAR 9.505(a). There are multiple such organizational conflicts of interest (“OCIs”) here. First, “biased ground rules” can result in skewed requirements, whether intentionally or not, in favor of consultants that are drafting requirements for procurements in which they later may be involved. Second, the dual roles of CH2M Hill and Battelle also create a concern that these consultants, by virtue of access to and knowledge of agency deliberations and plans, will have an unfair advantage in any competition for later work. Third, these circumstances give rise to an “impaired objectivity” OCI, because CH2M Hill and Battelle are apparently evaluating their own deliverables, *i.e.*, the prior Draft Reports, as the basis for recommending the Draft Parcel G Work

Plan for themselves. Because CH2M Hill and Battelle, and others, served as the consultants creating the Draft Reports, their ability to render impartial advice to the Government is completely undermined by their dual roles in drafting the supposedly “independent” Parcel G Work Plan.

In sum, having worked at the HPNS Site for many years, TtEC understands the challenges faced by the Navy (as well as EPA), given the complexity of the Site and the stakeholders involved. At this juncture, TtEC respectfully submits that it is imperative that the parties develop a process to fairly review the contractual work, as well as any additional efforts that the Navy (or EPA) may want to consider.

Given the importance of HPNS, including past work performed under multiple award contracts by TtEC and other contractors, determining an appropriate approach to Parcel G will be key to maintaining consistent standards for the Site. But, as it stands, the technical and programmatic deficiencies in the Parcel G Work Plan will only confuse the issues by applying standards that do not correlate to the contractual work or with the health-based standards set forth in the Parcel G Record of Decision. TtEC is prepared to commit its resources, including longstanding historical knowledge of the Site, to reviewing the attached Technical Comments with the Navy, in detail, and to discuss revising the Parcel G Work Plan to create a technically sound and comprehensive approach to the Site.

Sincerely,

A handwritten signature in black ink, appearing to read 'A.N. Bolt', written in a cursive style.

A.N. Bolt, PE, PMP  
President, Tetra Tech EC, Inc.

Enclosure

cc: Angeles Herrera, U.S. Environmental Protection Agency  
Janet Naito, California Department of Toxic Substances Control  
Anthony Chu, California Department of Public Health  
Greg Wagner, San Francisco Department of Public Health  
Nadia Sesay, Office of Community Investment and Infrastructure  
Terry Seward, Regional Water Quality Control Board

**Technical Comments on Hunters Point Naval Shipyard  
Parcel G Removal Site Evaluation Work Plan (June 2018)  
Submitted by Tetra Tech EC, Inc.**

The Navy's Parcel G Removal Site Evaluation Work Plan (Work Plan) must be revised to address serious methodological deficiencies. Specifically:

1. The Work Plan must clearly state the goals of the proposed investigation and remediation at Parcel G, and must revise the flawed Conceptual Site Model and Data Quality Objectives on which the proposed scope of work is based.
2. The Work Plan should not rely on the incorrect and unscientific conclusions in the Navy's review of previously collected data at the Site.
3. The Navy should revise the inadequate background sampling plan.

The Navy should correct these errors and revise the Work Plan to develop a scope of work that is appropriate for addressing the question that the Work Plan is intended to answer—*i.e.*, whether the remediation of Parcel G has satisfied the health-based standards set forth in the Parcel G Record of Decision (ROD). A properly optimized survey design would (1) take into account the prior investigation and remediation work that has been performed on Parcel G; (2) conduct an objective analysis of the validity of data that was collected during the Site remediation and investigation; and (3) collect verification samples in localized areas where additional sampling is determined to be necessary to confirm compliance with the Parcel G ROD. This approach will allow the Navy to achieve the Work Plan's goals in an efficient, reliable manner, with a degree of accuracy, integrity, and scientific rigor that is absent from the current Draft Work Plan.

Tetra Tech EC, Inc.'s (TtEC's) comments on the Work Plan are set forth in detail below.

**1. The Work Plan Lacks a Defensible Basis and Is Unlikely to Achieve Its Stated Objectives.**

Any plan for site investigation or remediation rests on three pillars:

1. The goals to be achieved;
2. The Conceptual Site Model (CSM); and
3. The outputs of the Data Quality Objectives (DQO) process.

The DQO process formally incorporates the goals and the CSM within an iterative design process that should be informed by the needs of all stakeholders, the site data, and science.

### **1.1 The Work Plan's goals are not sufficiently well defined or precisely described.**

The Work Plan states that its purpose is “to determine whether current site conditions are compliant with the remedial action objective (RAO) in the Parcel G Record of Decision (ROD).” [Work Plan at pp iii and 1-1]. The RAO for “radiologically impacted soil and structures” is to “prevent exposure to radionuclides of concern in concentrations that exceed remediation goals for all potentially complete exposure pathways.” [Navy 2009, at p. 29]. The ROD’s Table 5 lists numerical values of these “remediation goals” (RGs). [*Id.*, at p. 31].

EPA's comments on the Work Plan [EPA 2018] incorrectly state that the Parcel G ROD requires a point-by-point comparison of every sample measurement to the RGs set forth in the ROD. This is incorrect. Instead, the ROD requires the Navy to remediate radiological contaminants on Parcel G to achieve the risk-based standards set forth in OSWER Directive No. 9200.4-18 [EPA 1997] and/or the standard for unrestricted use set forth 10 CFR Section 20.1402 [EPA 2009, at p. 31]. Nothing in these standards requires a point-by-point comparison of concentrations to RGs. Such a comparison is an improper application of the risk-based decision-making process set forth in EPA's guidance [US EPA 1997].

In fact, the numerical values of the RGs have meaning only in the context of the “exposure pathways,” where it is understood they are related to estimates of risk to human health or the environment. The soil RGs for all radionuclides of concern (RoCs), except <sup>226</sup>Ra, were explicitly “derived from the EPA preliminary remediation goals (PRGs) based on an increased lifetime cancer risk range of 10<sup>-6</sup> to 10<sup>-4</sup> for future use scenarios.” [Navy 2008, at p. 3-2].

“Lifetime cancer risk” is estimated from *total* expected lifetime exposures. In no case will the total exposure for any person correspond to the concentration of a single particle of soil, nor will it correspond to the largest concentration that person ever encounters. A valid exposure scenario therefore incorporates assumptions about the *spatial and temporal extents* of the site materials any person might routinely encounter. Although those assumptions do not appear explicitly in the ROD, they are incorporated through its explicit references to exposure pathways and cancer risk.

The Work Plan does not acknowledge or reflect how the RGs are based on a risk assessment. This omission has serious implications for the investigation, as will be explained further below.

## **1.2 The Conceptual Site Model does not establish the need for further investigation.**

Section 2 of the Work Plan outlines the CSM in Table 2-1 [pp. 2-2 through 2-4]. Table 2-1 relies on the 2004 Historical Radiological Assessment (HRA) [Navy 2004] and acknowledges some of the extensive remedial work performed since that time, including:

- More than 4 miles of trench lines and 50,000 cubic yards of soil were investigated and disposed or cleared for use as onsite fill.
- Trench excavations were backfilled with homogenized clean, tested soil from onsite fill, offsite fill, or a mixture of both.
- Contaminated sinks and drain lines were removed from Room 47 of Building 351A.
- Liquid waste tanks, soil contamination, and contaminated sinks and drain lines were identified and removed from the Buildings 317/364/365 site.
- Sanitary sewers and storm drains and at least 1 foot of soil surrounding the pipes were removed. The sewer lines were removed to within 10 feet of all buildings. Impacted buildings had all remaining lines removed during surveys of the buildings. Non-impacted buildings had surveys performed at the ends of pipes and the pipes were capped.

Because of this work, all sources of radiological contamination identified in the HRA have been removed. Consequently, the Work Plan acknowledges there is a lower potential for radiological contamination at the Site than what was described in the original 2008 CSM [Work Plan at p. 2-4]. The Work Plan cites data collected at the Site to support this conclusion [*Id.*]

However, the Work Plan's CSM notes "Uncertainties" [p. 2-4] that the Navy appears to rely on to justify re-doing a significant portion of the work performed since 2004 on Parcel G, including (1) the "potential for data manipulation or falsification," (2) "data quality deficiencies," and (3) the presence of trenches "where scan data exceeded the investigation level and biased soil samples were not collected."

With respect to the first "uncertainty," it is critical to distinguish between what the Navy characterizes as "potential" data falsification and the actual, proven data falsification that has been identified at the Site. The latter was carried out by a small group of individuals, two of whom have been convicted of crimes based on their admitted misconduct. The Navy and Tetra



Tech EC, Inc. (TtEC) responded aggressively to the discovery of this misconduct by re-investigating and further remediating the affected areas [TtEC 2014].

As for the second and third “uncertainties,” these supposed justifications for re-doing the work on Parcel G result from unproven allegations reported in draft, unpublished reviews conducted by CH2MHill and other private consultants for the Navy [Navy 2017a-d; Navy 2018]. General and implausible allegations of “potential” data falsification are not a valid basis for a Work Plan. Instead, analysis is needed to characterize the implications of these suspicions. As later comments will demonstrate, in many circumstances the alleged problems with data quality or reliability are benign, because—even if they turn out to exist—they did not affect the decisions made during the investigation and remediation of Parcel G. In many circumstances the alleged problems pertain to the form and extent of documentation of the investigation and not to the results of the investigation. In other cases, the allegations stem from an incomplete review of the documentation or from a lack of understanding of how the investigation of Parcel G was carried out. Such issues of interpretation and documentation ought to be addressed by investigating the full record of the Parcel G investigation rather than by additional investigation of Parcel G (with its attendant costs in time and resources).

EPA's comments on the Work Plan [EPA 2018] similarly fail to account for the vast amount of investigation and remediation work that has been performed at Parcel G. By reverting to the 2008 CSM [EPA 2018, at p. 1], which was prepared before the completion of the remedy set forth in the Parcel G ROD, the EPA comments fail to account for work done at the Site since 2008. This is contrary to EPA's own guidance, which requires EPA to consider all “available data that have previously been collected for a site” in designing a remedial investigation to “avoid duplication of previous efforts and lead to a remedial investigation that is more focused and, therefore, more efficient in its expenditure of resources” [EPA 1988, at p. 2-5].

### **1.3 The Data Quality Objectives process lacks essential inputs and outputs.**

The Multi-Agency Radiation Survey and Site Investigation Manual (“MARSSIM”) [US EPA *et al.* 2000] strongly recommends using the Data Quality Objectives process to conduct the planning phase of a survey design, especially for complex investigations. The DQO process is an iterative procedure, formalized by the EPA in the 1980s and incorporated within MARSSIM, to “improve the survey effectiveness and efficiency and thereby the defensibility of decisions” [*id.* at p. D-1]. The process focuses on three elements: (1) **making decisions** (2) based on **observational data** about (3) **site conditions**.

1. **Decisions** are recommendations for action, such as to continue investigation, perform remediation, or approve a portion of the site for release. The basic unit of decision-making at this site is the “survey unit,” or SU. Building interiors have been divided into individual SUs. Areas of trench investigation have been divided into SUs known as “Trench Units,” or TUs. Decisions about the disposition of soil removed from TUs were based on temporary plats of soil called “(excavated) fill units.”
2. **Observational data** are quantitative measurements of an SU or of materials derived from it. They may include measurements of samples, readings from radiation scanners, and even narrative descriptions.
3. **Site conditions** are the *true* locations, amounts, and relevant physico-chemical properties of all RoCs within any SU. “True” is used to distinguish site conditions from *measurements* or *inferred conditions*, which are expected to approximate the true conditions but will unavoidably differ from them to some extent.

Decisions are framed in terms of *decision rules*. A decision rule unambiguously describes a set of possible site conditions and states what decision would be made when such conditions hold. However, site conditions are never known with perfect certainty. They must be deduced through scientific and statistical inference from observational data. There will always exist some chance of making an incorrect decision because the *data* or the *inferential procedure* incorrectly characterizes the site conditions. For example, all measurements of RoCs in soil samples at an SU might be less than the Remedial Goals, implying the SU can be cleared for release; yet there is always a possibility that some small amount of the SU soils with contamination have been overlooked.

The DQO process recognizes that such *decision errors* are unavoidable. The process works to limit them to rates that are satisfactory to all entities that might be affected by the decisions: the so-called *stakeholders*. Accordingly, it is essential that all stakeholders be involved in the DQO process [US EPA *et al.* 2000, at p. 3-2].

Section 3.1 of the Work Plan presents the results of an alleged DQO process, but it shortens or corrupts many of the requisite steps, described below.

- **State the Problem (Step 1).** MARSSIM asserts: “Since many sites or facilities present a complex interaction of technical, economic, social, and political factors, the success of a project is critically linked to a complete but uncomplicated definition of the problem.” [US EPA *et al.* 2000, at p. D-4]. This requires the decision-makers to “identify members of the planning team and stakeholders” [id.] and involve them in developing the DQOs.

The Work Plan provides no evidence that obvious stakeholders (such as the public, land developers, or TtEC, whose work is explicitly questioned in the DQOs) were identified or consulted in developing the proposed scope of work.

- **Identify the Inputs to the Decision (Step 3).** The Work Plan is motivated by the “uncertainties” concerning “potential for data falsification” and “data quality deficiencies” [Work Plan at p. 2-4], but it makes no effort to identify existing data that can help resolve these uncertainties. Instead, the Work Plan implicitly assumes, without justification, that the only tool available to achieve its goals is further investigation of the Site.
- **Develop Decision Rules (Step 5).** The decision rules in the Work Plan are not clearly defined. To be actionable, the rules need to describe clearly, quantitatively, and in detail what it means for “site conditions [to be] compliant with the Parcel G RAO” [Work Plan at p. 3-1].
  - The rules are incomplete. They cover the cases where “site conditions are compliant with the Parcel G RAO” and otherwise where site conditions “exceed background levels.” No rule is provided for the case where site conditions do not comply with the RAO but also do not exceed background levels.
  - The rules contradict the stated objective (in Step 2), which is “to determine whether site conditions are compliant with the Parcel G ROD RAO.” The second rule replaces that with a vague goal of being “protective of human health.”

EPA's comments on the Draft Work Plan correctly identify the deficiencies in the decision rules proposed by the Navy [EPA 2018, at p. 6], but the rule proposed by EPA is incorrect and not based on sound science or risk assessment principles. A point-by-point comparison with ROD RGs is not required to achieve compliance with the health-based standards set forth in the ROD. A correctly implemented DQO process will identify and characterize those standards in a fashion that supports the quantitative analysis (in Step 7) that leads to designing an effective work plan.

- **Specify Limits on Decision Errors (Step 6).** The Work Plan replaces Step 6 of the DQO development process by a step titled “specify the performance criteria.” The Work Plan does not, however, provide these criteria. In their place, the Work Plan lists one statistical procedure for implementing one decision rule. The procedure compares all measurements of “all samples” of an SU to the numerical values of their RGs. This procedure is offered without statistical or scientific justification. No representations

concerning the decision error rates are made, and no decision error rates are even proposed. This is the step in the DQO process that depends the most on inputs from the stakeholders. It can be executed properly only by identifying the stakeholders and involving them in a legitimate DQO process in which they are given enough information to identify potential decision errors and propose appropriate limits on them.

- **Optimize the Design for Collecting Data (Step 7).** Step 7 is the culmination of the DQO process. Optimization is carried out by “formulating the mathematical expressions needed to solve the design problem for each data collection design alternative” and then “selecting the optimal design that satisfies the DQOs for each [such] alternative” [US EPA *et al.* 2000, at p. D-28]. The Work Plan provides no evidence that such a design problem was ever formulated, much less solved. Consequently, it lacks the foundation to establish that its procedures are optimal or even adequate.

Because the Work Plan fails to follow the DQO development process, it proposes a scope of work that cannot answer the question the Work Plan is intended to address—*i.e.*, whether the remediation of Parcel G achieved the health-based standards set forth in the Parcel G ROD. The Work Plan proposes to remove the durable cover (including asphalt, asphalt base course, concrete, gravel, debris, and obstacles that have been added since the original work was performed) in areas that have been previously remediated and to re-excavate and over-excavate previously remediated areas. The re-excavation and over-excavation of the trench units does little to advance the goal of preventing exposure to RoCs (and potentially increases the potential for such exposure).

A properly designed investigation of Parcel G would likely collect verification samples of materials beneath durable cover and in previously remediated areas through borings or localized trenching. This would allow for much faster and more efficient verification of the previous remediation work and would result in higher-quality data to give all stakeholders the opportunity to make decisions with known, mutually acceptable levels of confidence.

## **2. The Work Plan Relies on Flawed Analyses of Previously Collected Data.**

As noted above, the Navy and TtEC identified evidence of discrete and localized data falsification at the Site in 2012 and responded aggressively by re-investigating and further remediating the affected areas [TtEC 2014]. Following this investigation, former employees of a subcontractor have made general allegations of “manipulation” and “falsification” of data at the Site beyond the activities investigated and corrected by TtEC and the Navy.

The Navy assembled a technical team of Battelle, Cabrera Services, CH2MHill, Perma-Fix Environmental Services, and SC&A Environmental Services and Consulting (“reviewers”) to “conduct an evaluation of the previous data in light of the claims made” [Navy 2017a, at p. ii]. The reviewers’ work consisted of “assess[ing] the potential for data falsification or manipulation and recommend[ing] follow-up data collection to validate previous decisions regarding the property condition” [*id.*]. The reviewers’ findings are set forth in a series of data evaluation reports [Navy 2017a-d; Navy 2018], including a soils report for Parcels B and G [Navy 2017a] and a buildings report that addresses building scan data across the Site, including on Parcel G [Navy 2018]. These reports are in draft form and show no evidence of formal review or approval by the Navy or any regulatory agency.

## **2.1 The reviewers’ draft evaluation of building survey data does not support the Navy’s proposed investigation of Parcel G.**

The reviewers’ evaluation of the building survey data identified “duplicated” alpha/beta or gamma scan data in 24 of 220 survey units in Parcel G [Navy 2018, at Table 6-1]. The buildings report also concluded that average speeds for alpha/beta scans as estimated from selected data were 2.85 cm/s, “more than twice” 1.37 cm/s, the speed initially specified in the project work plan [*id.*, at p. 4-3] and that efforts to assess whether “the survey instrumentation was not in motion” during parts of certain surveys were “inconclusive” [*id.*, at pp. 5-1 and 5-4]. The draft buildings report confesses an inability to carry out a full analysis “because some types of data manipulation are difficult or impossible to identify” and therefore “additional data will need to be collected to support a recommendation for unrestricted radiological release” [*id.*, at p. 8-1].

Notably, the reviewers’ finding regarding building scan speeds is unrelated to any potential data falsification. Rather, the finding reflects the practical difficulties of moving scanners extremely slowly. The Navy itself approved faster scan speeds at the Site because, although they raise the limits of detection, the higher limits were still adequate to achieve the investigation objectives. Issues with scan speed were subsequently addressed by TtEC and corrected with the Navy’s concurrence. This finding does not trigger a need for a new investigation, nor does the inconclusive finding regarding whether survey instruments were stationary when data were being collected.

Only the reviewers’ finding regarding data duplication identified evidence of potential data falsification. Many of the conclusions the reviewers draw from that evidence are incorrect. The draft report documents the existence of data that appear twice in a database or data that have mistakenly been uploaded twice and attributed to different buildings or survey units. Such problems resulted from an approved work flow that required the data from each scanning

instrument to be manually edited in the field to include information that is not available from the instrument itself (including the survey location, type of material scanned, date of the survey, and identifier of the background dataset to use for reference).

These types of editing errors are routine in any large database—finding them is not evidence of falsification. For example, the buildings report identifies a “possible duplicate import” in footnote 7 to Table 6.2 [Navy 2018, at p. 6-6]. Many of the other duplicates are of the same nature but were not explicitly noted. Nevertheless, the buildings report indiscriminately lists all such instances in its summary and in the frequency statistics it reports [Navy 2018, at Table 6.1, pp 6-3 to 6-4]. These statistics do not reflect the extent of potential or actual “data falsification” or “data manipulation”; they only reflect the extent of data duplication.

Such errors often can be corrected by referring to original materials such as raw instrument logs, field log books, and project reports.

## **2.2 The reviewers’ draft evaluation of soil data does not support the Navy’s proposed investigation of Parcel G.**

The reviewers’ evaluation of trench, fill, and building soil survey units in Parcel G claimed to have identified evidence of “potential data manipulation or falsification” in 20 of the 63 trench units, 54 of 107 fill units, and 25 of 32 building soil survey units [Navy 2017a, at pp. iv-v]. Additionally, the soils report relates that “because it is impossible to determine whether every instance of potential data manipulation or falsification has been identified, the Navy recommends additional surveys and sampling beyond the areas with evidence of data manipulation” [*id.*, at p. 1-2].

The reviewers’ conclusions are based on “logic tests,” interpretations of statistical graphics, statistical tests, and judgments documented in Appendices A, B, and C of the report. As further explained below, these tests have little relevance to the question of potential data falsification, were poorly executed, and ignored alternative explanations of their results.

### **2.2.1 Logic tests**

Five “logic tests” examine the recorded dates of soil sample collection and analysis. A sixth logic test compares sample masses reported by the onsite and offsite laboratories. For example, the date comparisons call into question (or “flag”) sets of samples that were not all collected on the same day, or not all analyzed within a space of two working days falling no later than two weeks after sample collection. These are frequent occurrences arising from logistical constraints in performing the sampling and scheduling laboratory analyses. The date

comparisons also (rarely) identify true inconsistencies in the documentation, such as the appearance of measuring a sample before it was collected. Such inconsistencies can be resolved by consulting the field and laboratory records. The weight comparisons have similarly benign explanations, such as sending two separate physical aliquots of a sample to the two laboratories.

None of these flags indicate or are even related to data falsification. Indeed, they arguably should have the opposite interpretation: if an individual knowingly falsifies a chain of custody form, they likely will make an effort to *eliminate* any such inconsistency. Flagging database quality issues that are certain to arise routinely only hampers the review with a flood of irrelevant information. Thus, the reviewers' logic tests do not (and could not) provide evidence of data falsification that would support the resampling and re-excavation of Parcel G.

### **2.2.2 Statistical tests**

#### **(a) Statistical calculations in the buildings report are incorrect and misinterpreted.**

The buildings report performs four separate analyses. One of them, the "Data Distribution Comparison Method," relies on statistical testing. It employs the Kolmogorov-Smirnov Test ("K-S Test") [Navy 2018 at Section 5]. To carry out the test the reviewers use add-in software for Excel called "RealStats" [*id.* at p. 5-2]. This software computes "critical values" and "p values" to assess differences between pairs of datasets. The Data Distribution Comparison Method applies the K-S Test to alpha and beta results expressed in counts per minute (cpm). The results of this analysis are incorrect for the following reasons.

1. The RealStats calculations are valid only for "sufficiently large" sample sizes [RealStats 2018]. The datasets to which these were applied in the buildings reports are relatively small. These are the ones that matter, because the K-S Test will find even negligible differences in larger datasets to be "significant."
2. The RealStats calculations assume the data are drawn from "continuous distributions." The indication of a non-continuous distribution is the recurrence of one or more values in the datasets: a "tie." Because alpha and beta readings are relatively low counts (often less than 10 for alpha and in the hundreds for beta), ties are frequent.

For example, when—as a check—the K-S test is correctly computed for the only pair of datasets reproduced *in toto* in the buildings report [beta scans for Building 146 SU 33 walls, *id.* at p. 5-3], it yields a p-value of less than 0.01% (one part per ten thousand), indicating a significant

difference. However, the RealStats p-value shown in the buildings report is 9.3% [visible in Figure 5-1, *id.* at p. 5-3], which would *not* be taken as a evidence of a difference. This is a very large error in the report's analysis.

The buildings report also misinterprets the K-S Test results. The foregoing example compares an original set of static survey results to a later survey of the same area. The buildings report concludes: “In this case, the allegation [of a stationary detector during the original scan] could be true” [at p. 5-4]. However, plots of the data clearly indicate the verification scan exhibits the same spread in the detectable range of the data but has a slightly yet consistently lower response overall (by about 20 cpm). This is evidence of agreement between the original and verification scans and confirms that the work was carried out in accordance with the applicable work plan and contractual requirements, which provided for an initial, conservative scan followed by a more accurate verification scan (which showed a lower level of contamination). These conclusions thoroughly contradict the conclusion in the buildings report.

Because the buildings report uses flawed, inapplicable software and neglects to follow up its results by examining the data and interpreting its results, none of its conclusions from the Data Duplication Comparison Method is reliable.

**(b) The statistical tests in the soils report suffer from fundamental errors in their selection, application, and interpretation.**

The use of statistical tests lends the appearance of objectivity to the soils report. However, the selection, application, and interpretation of these tests are uniformly questionable for the following reasons.

1. Some of the tests are inappropriately used. For example, Appendix A to the soils report relies extensively on the Kolmogorov-Smirnov Test (“KS Test”) to identify “differences” among groups of soil measurements. This test will identify *any* kind of difference, not just a difference that might be evidence of data falsification. In particular, the variability of the natural geological materials and the artificial fill used to build up the Site causes the KS Test to identify “multiple populations” or differences so frequently that the authors of the soils report usually have to dismiss its results.
2. Some of the tests are inapplicable. For example, Appendix A to the soils report applies two versions of a “Benford Test” to each of 21 overlapping groups of data. Benford’s Law rests on the empirical observation that the initial digits appearing in many lists of numbers are not uniformly distributed, provided the numbers range over many orders of magnitude. It is most often used to flag unusual entries in financial documents.



Radiation measurements are not dollars. In these data, they never span the wide range of values needed for Benford's Law to hold. Predictably, in 100% of all applications of this test in the soils report, the data "failed." These results are meaningless. They serve only to create an impression of problems where none exist.

3. Some "tests" are not tests at all. One of the "tests" employed by the soils report is "hierarchical clustering," performed 64 times in Appendix A. Many subjective choices must be made to perform clustering analyses. The results often depend greatly on these choices. Because of this subjectivity, and because the usual apparatus of statistical testing is absent (a null hypothesis, a test statistic, and a p-value), cluster analysis is not a formal "statistical test." The soil report misleads readers by characterizing it as a "statistical test" and failing to explain its exploratory, subjective nature.
4. Some of the test results are incorrectly calculated. For example, Appendix A includes plots showing thousands of "confidence intervals" for average activities of radionuclides. Those plots purport to indicate which survey units have "significantly high" or "significantly low" activities. This determination is crucial to the data evaluation. However, those confidence intervals are incorrect in two ways.
  - a. The average of data from each survey unit is compared to a confidence interval that is computed from *all* data within the parcel, including that survey unit. It is invalid to compare confidence intervals directly, and it is also invalid to compare confidence intervals that are based on overlapping data as if they were statistically independent.
  - b. The plots in Appendix A show confidence intervals associated with survey units that have just a single measurement. Confidence intervals reflect *variation* in data. No variation can be observed in just one number. There are no standard statistical methods to compute confidence intervals based on single numbers. Either incorrect calculations were performed or strong, unstated assumptions were adopted. This problem suggests systematic methodological errors that call into question all of the confidence intervals calculated by the reviewers.
5. Some of the tests are incorrectly interpreted. An example of such an error is the procedure to flag "outlying" data. The Data Evaluation Forms in Appendix C frequently refer to "high outliers" and "low outliers." These are labels attached by the analyst to data that seem extreme. The graphics in Appendix A often identify outlying data. For instance, bar charts of daily mean results graphically distinguish dates where the mean

is “significantly higher than other days” or “significantly lower than other days.” This form of flagging suffers from several problems:

- a. The methodology is inconsistent. In some cases there are unflagged daily average measurements that are *more* extreme than so-called outliers.
- b. The methodology neglects correlations among the data. The measurement of any sample consists of many radionuclide activities, not just one. A true “outlier” ought to be assessed in this multivariate context.
- c. The methodology flags too many values. Outliers *must* be rare; otherwise—practically by definition—they not outliers.

Another example is the misuse of confidence intervals. Appendix B of the Work Plan states:

*A “Units Evaluation Flag” was applied to a survey unit data set if the mean concentration of results for a radionuclide from a unit falls outside of the 95 percent confidence interval of the mean concentration of results for that radionuclide in a given parcel.*

[Appendix B at p. 1.] A 95% confidence interval, by definition, has a 95% chance of including the true but unknown mean of a population that was randomly sampled. As such, it is valid to use a confidence interval to make decisions about that mean, but not to compare it to random quantities such as the average of a set of data. (For that purpose, a t-test or a prediction interval may be used, among other alternatives.) The failure of a random quantity like the “mean concentration of results” to lie within a confidence interval may only reflect chance variation due to the sampling and measurement process. Thus, this misuse of confidence intervals will tend to flag data incorrectly.

6. The test results are misleading. The statistical test results need to be adjusted for multiple comparisons and correlations. The soils report presents the results of tens of thousands of tests and applies each test repeatedly to subsets of the same data (*e.g.*, all data in a survey unit; only the data produced by the onsite lab; only the data produced by the offsite lab; all data aggregated into the parcel). The report also applies tests to each radionuclide separately, even though their measurements are expected to be correlated. The execution of many interrelated statistical tests in the pursuit of “significant” results is widely but disparagingly known as “data dredging,” “data

snooping,” “p-hacking,” or a “fishing expedition.” It has been politely described as “a willingness to look hard for patterns and report any comparisons that happen to be statistically significant” [Gelman and Loken 2013]. Data dredging results in invalid statistical inferences because, through the sheer number of tests that are performed, it is guaranteed to produce some results that are nominally “significant.” In the context of extensive data dredging, no individual statistical test result can be considered significant.

7. Statistical testing is performed inconsistently. The narrative portions of the Data Evaluation Forms (Appendix C) selectively reject or ignore the results of many of the tests that are performed. In the sections on “additional database review” and “adjacent survey/trench review,” though, quantitative comparisons of soils data are frequently undertaken without employing any statistical test at all. These unexplained choices of when to use statistical procedures and when not to makes the results appear subjective.
8. The test procedures are not documented. The interpretation of statistical tests often rests on small details about their underlying assumptions and how the calculations were carried out. These details are sufficiently important that leading scientific societies and journals establish standards for reporting test results and will reject work that does not meet those standards. The soils report (including its appendices) has no adequate descriptions of any of the tests it applies. As such, it does not meet minimal standards of scientific or statistical communication.

**(c) The logic of statistical testing evidenced in the soils and buildings reports is invalid.**

The US EPA claims to have found “potential signs that 90 to 97% [of previous data] were unreliable” [US EPA 2018]. This belief appears to derive from an internal evaluation that was conducted with methods and strategies akin to, or possibly identical to, those used in the draft reports. It is not possible to evaluate the validity of the EPA claim in detail because the EPA has not adequately documented its review. Evidence found in a summary of the results [US EPA 2017] indicates the claim is derived from, and suffers from the same methodological problems exhibited by, the draft soils report [Navy 2017a].

Independent of these problems in conception and execution, the underlying strategy and logic of the logical and statistical tests are wrong. The tests seek evidence of data falsification or manipulation in the form of patterns that might (hypothetically) arise from such activities *as well as occurring routinely for many benign reasons*. As such the approach is indiscriminate and

unscientific. This is already evident in the minutes of the “scoping meeting” convened by the Navy in 2016 to begin the radiological data evaluation [Navy 2018b, Attachment 1]:

*Further concerns were expressed regarding the data that does not show any obvious anomalies. It is her [Lily Lee, US EPA] opinion that since Tetra Tech has disclosed that data has been falsified, we cannot say that the data is reliable even though the statistical tests do not turn up any results. Scott Hay [Cabrera Services] and Bob Kirkbright [CH2M] explained that our statistical tests will identify anomalies in the data, including running tests designed to identify instances where data may have been falsified . . . .*

*Scott Hay explained we will be using a test on the data sets where problems have already been identified, as well as the data set in its entirety. If these tests are able to identify the known problem areas, it will provide confidence in the analysis.*

[*Ibid.* pp 2-3.] An approach that is fair, practical, and scientific would focus on identifying areas of the Site that do *not* require re-investigation. Instead, the Navy and EPA reviews were designed to have high *sensitivity* (ability to identify falsified datasets) but were unconcerned about *specificity* (ability to correctly identify reliable data). Sensitivity and specificity are about decision errors, an issue not properly addressed in the draft Work Plan (*q.v.*).

A correctly performed statistical review of the Parcel G soil samples would provide meaningful information that could be used to determine whether and to what extent further investigation of the Parcel might be necessary. However, the flawed methodology and errors in the soils report make it (and all derived or related work such as the EPA’s review) unreliable for this purpose.

For these reasons, the Navy should not rely on the draft soils report or on the EPA’s claims as a basis for additional investigation and remediation of Parcel G.

### **2.2.3 Graphical assessment**

Graphics are the heart of the Parcel G soils report [*see, e.g.,* Appendix A]. They comprise over 12,000 of its nearly 14,000 pages. They display the data on which the report depends. They furnish the material that initially suggests—both to the report’s authors and its readers—what course of data evaluation to pursue, which hypotheses about data falsification to entertain, and how to evaluate those hypotheses. Accurate representation of the data is fundamentally important.

The graphics in the soils report uniformly distort and misrepresent the data, for the following reasons:

1. The graphics use distorted axes.
  - a. The many hundreds of overlapping histograms in Appendix A, used to compare datasets, implicitly employ two scales that differ by up to a factor of five but display the values of only one scale. **This deceives the viewer and exaggerates the differences.**
  - b. “Time series plots” in the Data Evaluation Forms of Appendix C show data collected at different times as if they were spaced evenly throughout. **This distorts trends and provides an inaccurate picture of the course of the investigation.**
  - c. Side-by-side box plots in the Data Evaluation Forms of Appendix C, used to track changes in data during the investigation of each trench unit, vary their scales and their origins. **This exaggerates, suppresses, or even reverses apparent trends.**
  - d. “Probability plots” in the Data Evaluation Forms of Appendix C usually compress their value scales. **This frequently makes it impossible to interpret the plots at all.**
2. Bar charts, which purport to display individual data values, systematically obscure the smallest values. This distortion creates an overwhelming visual impression of much greater radioactivity at the Site than really existed.
3. Some graphics obscure details by degrading graphical quality. The Data Evaluation Forms of Appendix C routinely rely on three sets of graphics, comprising almost half of Appendix C. These graphics are reproduced at such low resolution that details are impossible to read in most of them.
4. Some graphical designs are ineffective. The bar charts used to display individual data values tend to obscure the data. A better choice, which is widely available and easily understood, is to create plots that mark data with point symbols rather than bars. In addition, most of the graphics show only one variable or one variable in a time context. Multivariate graphics (such as scatterplot matrices and other “small multiples” [Tufte 1983]), which reveal relationships among two or more variables, are entirely absent.

5. The sheer quantity of graphics hinders their use. The 12,000 pages of Appendix A could be reduced to a few hundred pages by means of suitably chosen multivariate graphics, small multiples, and applying Edward Tufte's design principle of *minimizing the data:ink ratio* [Tufte 1983]. This is done not by shrinking the plots, but by eliminating redundant and superfluous graphical elements, choosing effective methods of graphical representation, and organizing the plots in systematic and meaningful ways. Thousands of pages of plots overwhelm, whereas a few well-designed pages can *inform* and *illuminate*.

Collectively, the graphics produce an "illusory truth" effect. Repetition of a false claim causes even smart, educated people to increase their belief in its objective truth [Hasher *et al.* 1977]. The claim need not be very plausible for this to occur [Pennycook *et al.* 2018]. A distorted graphic is a false claim, even when accompanied by a correct account of the data it presents [Tufte 1983 at p. 77]. The extensive repetition of distorted, inadequate, poorly chosen, or unreadable graphics is likely to cause readers of the soils report to be convinced of conclusions that are not supported by the data.

#### **2.2.4 Data interpretation**

The interpretations of data in the soils report [Navy 2017a] frequently neglect contextual information that points to alternative conclusions. Examples of these lapses include:

1. Failure to account for laboratory qualifiers. Both the onsite and offsite laboratories associate "qualifiers" with the numerical measurements they make. Qualifiers indicate which measurements should be considered indistinguishable from zero ("nondetects"), which ones are less precise than usual, and which ones ought to be rejected out of hand. For instance, the soils report flagged data when "sample results for naturally occurring radionuclides were at or below zero" [Navy 2017a, at p. 3-3] even though this frequently occurs because values are below the limits of detectability, have had positive background values subtracted from them, *and have been qualified as imprecise by the laboratory*. The graphical displays promote such misinterpretations by neglecting to distinguish qualified data from unqualified data.
2. Failure to account for subsequent investigation and remediation. The Navy and TtEC acknowledge that suspect data entered the database (through the unlawful actions of two individuals who later confessed to data falsification). However, TtEC and the Navy aggressively investigated this suspect data, as documented in [TtEC 2014], and conducted additional investigation under close Navy supervision. The soils report does

not distinguish suspect data as corrected by this investigation from other data that the reviewers have newly characterized as potentially falsified. The soils report (as well as the buildings reports) needs to make this essential distinction.

3. Reliance on apparent misunderstandings of the investigation protocols. The authors of the soils report do not appear to have an adequate understanding of the investigation protocols at Parcel G. As an example of a misunderstanding, the Data Evaluation Forms in Appendix C flag many survey units because of an alleged failure to investigate contamination. For instance, Appendix C states:

*The gamma scan of the trench (range of 3,700 to 7,400 cpm) reported in Attachment 1 Radiation/Contamination Survey Form of the SUPR exceeds the investigation level (7,048 cpm). This contradicts the statement in Section 3 of the SUPR [Survey Unit Project Report] that no scan results exceeded the investigation level. The survey activity apparently failed to respond to elevated gamma scan results.*

[Navy 2017 a, Appendix C at p. 105, discussing TU 77.] The SUPR documents the “Investigation Level,” a number based on a background scan made by the same instrument. It also reports the range of results observed during an initial gamma scan of the unit. (Individual results were not saved in the data logger, but some—including the smallest and largest—were recorded in the field by the surveyor.) These results were generated by a surveyor who moved slowly and systematically around each survey unit while monitoring the instrument for readings above the Investigation Level. To improve precision, the investigation protocol required the surveyor immediately to measure the location of any elevated reading for a longer duration. In many cases this follow-up was below the Investigation Level, establishing that the elevated reading resulted from statistical fluctuation. Such transient random events were expected to happen frequently. Contrary to the conclusions reached in the soils report, they reflect nothing about site conditions and were therefore not documented except in the form of the maximum reading recorded in the SUPR. Thus, the cases where the maximum exceeded the Investigation Level were ordinary occurrences rather than evidence of any data falsification.

4. Making invalid assumptions about correlations between scan and soils data. The soils report makes invalid assumptions about the correlations between scan and soils data. When the higher-precision follow-up measurement to a scan exceedance still indicates a value above the Investigation Level, the location of the scan is marked in the field and

scheduled for “biased” sampling. The soils report flags circumstances where the subsequent measurements of the biased samples were nevertheless not high enough to require remediation. An example is ES092 (fill unit 92):

*The SUPR reported elevated gamma scan measurements and the collection of biased samples; however, no remediation was performed. This narrative is consistent with the allegation that biased samples were collected in areas to avoid potentially elevated sample results.*

[Navy 2017a, Appendix C at p. 1190] Such flagging reflects the invalid assumptions that (a) gamma scan data are as reliable as laboratory measurements of soil samples and (b) there must be a perfect correspondence between gamma readings above the investigation level and contamination in soils. The gamma measurements are a *field screening* device. They are inherently less accurate and less precise than the laboratory measurements. An exceedance of the Investigation Level in the field will not invariably lead to soil measurements that exceed the Remedial Goals. It is erroneous to interpret such circumstances as even “potential” evidence of “potential” data falsification; rather, they are evidence that the protocols for field screening and laboratory testing of soil samples functioned as they should.

5. Misinterpreting inter-laboratory comparisons. The use of an onsite laboratory balanced the objectives of (1) making cleanup decisions with acceptable confidence and (2) permitting the investigation to proceed at a rapid pace. The onsite laboratory provided analytical results within days, whereas the offsite laboratory methods require more than three weeks (for sample shipment, testing, and reporting). This rapid turnaround enabled the investigation to respond rapidly and appropriately to the sample results and provided it the flexibility to take large numbers of measurements as needed. Such flexibility in measurement usually comes with lower accuracy. (The onsite laboratory nevertheless was lauded by the Navy for its high-quality work and approved by the US EPA.) The investigation design compensated for that possible loss of accuracy by working to assure the onsite laboratory results were unlikely to err on the low side (especially for <sup>226</sup>Ra). The price was the *predictable* need to remediate more soil than necessary. This was not an error—it was part of the remedial design approved by the Navy—nor was it uneconomical, because of the benefits it brought in the form of a faster and more flexible investigation.
6. Relying on untenable physical and statistical assumptions. Many incorrect physical and statistical assumptions are implicit in the soils report. For example:



- a. The “uniformity of remedial effects” theory. The report expresses this theory as follows:

*If remediation was performed for a radionuclide of concern (ROC) other than Ra-226, it was assumed that the concentrations of Ac-228, Bi-214, and K-40 would generally remain consistent for the entire data set.*

[Navy 2017a, Appendix B at p. 3] This is tantamount to assuming homogeneity of geological materials throughout any trench unit or fill unit. It is unsupported by the history of the site or geological facts. Rather than arbitrarily flagging a dataset for this reason, the reviewers should have investigated the geological constituents of the materials in and around the trench to determine their degree of homogeneity.

- b. The “two population” theory. This theory was expressed as follows:

*“An unusual distribution of final systematic sample results that is a potential indication of two or more data populations” is potential evidence of data falsification.*

[Navy 2017a, Appendix B at p. 4.] Again, this conflates geological heterogeneity with “potential evidence of data falsification.”

- c. The “limited variability” theory. The soils report states this theory as follows:

*“Normal statistical variation at the 99% confidence limit for a reading at the center of the span of these measurements would account for essentially the entire observed variation in the static measurements. The variation of the soil concentration of the primary gamma-emitting radionuclides that was observed in the laboratory results is therefore not reflected in the gamma static results.”*

[Navy 2017a, Appendix C at p. 434.] This passage responds to allegations that sometimes, instead of performing a survey, the surveyor would just set the instrument on the ground to collect a series of readings at one location. It vaguely suggests a statistical test to determine whether that occurred. However, even if this allegation were true, the measurements would not all be the same,

because rates of radioactive decay change randomly over time even in a fixed spot. The test relies on the (valid) idea that variability in any series of measurements can arise from multiple distinguishable sources: in this case, variability in the rate of radioactive decay (which is always present) and variability arising from other causes, such as differing amounts of radioactivity at different locations throughout the survey unit. The test needs to evaluate whether that component of *spatial* variability is present or absent. Computing “normal statistical variation at the 99% confidence limit,” whatever that might mean, does not isolate the spatial variability, does not correctly quantify it, and is not an appropriate way to carry out such a test.

### **3. The Proposed Background Sampling Plan Is Inadequate.**

The background sampling plan, presented in the Work Plan’s Appendix A, is unscientific, inadequate, and threatens to undermine the validity of the entire proposed Parcel G investigation.

- The background sampling plan does not adequately describe the procedures that will be used to determine background concentrations. The background sampling plan uses vague language, promising only to make “comparisons” between background and onsite results or to “define additional data sets” based on “visual observations” [Work Plan, Appendix A at 3-3].
- The background sampling plan’s Data Quality Objectives are incomplete and inadequate. The background sample plan’s “decision rules” address only minor technical issues rather than the principal decisions themselves. The plan has no effective performance criteria. The DQOs neither acknowledge nor use existing data—most of which were collected in the same areas and therefore will be useful both for designing the background investigation and optimizing its costs.
- The proposed sample design will not collect representative data. By focusing the sampling within small portions of the available background sampling areas, it is likely not to capture the full variation of background conditions.
- The statistical procedures have no scientific or mathematical justification. The proposed methods to characterize background are not suitable for the decisions that must be made concerning whether onsite samples would exceed background. When they are applied, they are certain to mischaracterize background. Errors in the statistical procedures used include:

- Automatic detection of “outliers,” which threatens to discard any high or low values that signal natural heterogeneity in background.
- “Determination of statistical differences between datasets,” which (even if correctly carried out) will erase all evidence of geological heterogeneity in background.
- Failure to distinguish conditions in surface soils from those in subsurface soils, which are likely to differ due to variation in geological materials and vertical transport of fallout.
- Failure to recognize the likelihood of high positive correlations among measurements made within a common boring.

Characterizing background correctly is crucial for any investigations of the Site itself, because an error in one direction can cause most Site data to appear “elevated” when, in fact, they are not, while an error in the other direction can mask elevations in all or most of the Site data wherever they might occur. The deficiencies in the background sampling plan will assure the failure of the entire investigation.

#### 4. References

Gelman, Andrew and Eric Loken, 2013. *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time.* Columbia University. Available online at [www.stat.columbia.edu/~gelman/research/unpublished/p\\_hacking.pdf](http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf) (accessed Aug. 25, 2018).

Hasher et al. 1977. Hasher, L., D. Goldstein and T. Toppino. *Frequency and the Conference of Referential Validity.* J. Verbal Learning and Verbal Behavior 16, 107-112 (1977).

Navy 2004. Naval Sea Systems Command. Historical Radiological Assessment, Volume II, Use of General Radioactive Materials, 1939-2003 (Aug. 31, 2004).

Navy 2008. Final Radiological Addendum to the Revised Feasibility Study for Parcel D, Hunters Point Naval Station (Apr. 11, 2008).

Navy 2009. Record of Decision for Parcel G, Hunters Point Shipyard, San Francisco, California (Feb. 18, 2009).

Navy 2014. NAVFAC Presentation to CDPH, RASO, DTSC, and BRAC PMO (Sept. 17, 2014).

Navy 2017a. Department of Navy, Naval Facilities Engineering Command, Base Realignment and Closure Program. Draft Radiological Data Evaluation Findings Report for Parcels B and G Soil (Sept. 2017).

Navy 2017b. Department of Navy, Naval Facilities Engineering Command, Base Realignment and Closure Program. Draft Radiological Data Evaluation Findings Report for Parcels D-2, UC-1, UC-2, and UC-3 Soil (Oct. 2017).

Navy 2017c. Department of Navy, Naval Facilities Engineering Command, Base Realignment and Closure Program. Draft Radiological Data Evaluation Findings Report for Parcel C Soil (Nov. 2017).

Navy 2017d. Department of Navy, Naval Facilities Engineering Command, Base Realignment and Closure Program. Draft Radiological Data Evaluation Findings Report for Parcel E Soil (Dec. 2017).

Navy 2018. Department of Navy, Naval Facilities Engineering Command, Base Realignment and Closure Program. Draft Building Radiation Survey Data Initial Evaluation Report (Mar. 2018).

Navy 2018b. Draft Parcel G Removal Site Evaluation Sampling and Analysis Plan Former Hunters Point Naval Shipyard San Francisco, California. August 2018

Pennycook *et al.* 2018. Pennycook, G., T.D. Cannon and D.G. Rand. *Prior exposure increases perceived accuracy of fake news*. J. of Experimental Psychology: General [in press].

RealStats Using Excel 2018. Two Sample Kolmogorov-Smirnov Test. Online at <http://www.real-statistics.com/non-parametric-tests/goodness-of-fit-tests/two-sample-kolmogorov-smirnov-test/>. Accessed Aug. 6, 2018.

TtEC 2014. Tetra Tech EC, Inc. Investigation Conclusions, Anomalous Soil Samples at Hunters Point Shipyard (Apr. 2014).

Tufte 1983. Edward Tufte. *The Visual Display of Quantitative Information*. Cheshire Press (1983).

US EPA 1988. US Environmental Protection Agency. Guidance for Conducting Remedial Investigations and Feasibility Studies Under CERCLA. OSWER Directive No. 9355.3-01 (Oct. 1988).

US EPA 1997. US Environmental Protection Agency. Establishment of Cleanup Levels for CERCLA Sites with Radioactive Contamination. OSWER Directive No. 9200-4-18 (Aug. 22, 1997).

US EPA 2017. EPA, DTSC, and CDPH reviews of the Navy's Draft Parcel G Radiological Data Evaluation Findings Report Draft Hunters Point Naval Shipyard, San Francisco, California. SEMS-RM DOCID # 100009183. Available online at <https://semspub.epa.gov/work/09/100009183.pdf> (Aug. 25, 2018).

US EPA 2018. Letter to L. Lansdale, Base Realignment and Closure Program Management Office, US Department of Navy, from Angles Herrera, US Environmental Protection Agency Region IX (Aug. 14, 2018).

US EPA *et al.* 2000. US Environmental Protection Agency *et al.* Multi-Agency Radiation Survey and Site Investigation Manual, NUREG-1575 (Aug. 2000), with Errata and Addenda (Aug. 2002).